Relátorio - Redes Neurais Artificiais
Desvendando Mistérios Médicos: Explorando a Fronteira da Detecção de Câncer de Mama com Redes Neurais

1st RUAN RODRIGUES SOUSA
Universidade Federal do Ceará
Campus Jardins de Anita
Curso de Ciência de Dados
Itapajé, CE-Brazil
ruanrodrigues@alu.ufc.brl

2st LARISSA VITÓRIA VASCONCELOS SOUSA
Universidade Federal do Ceará
Campus Jardins de Anita
Curso de Ciência de Dados
Itapajé, CE-Brazil
larissavvsousa@alu.ufc.br

***Abstract*—A análise abrangente do conjunto de dados *Breast Cancer Wisconsin* concentrou-se na distinção crucial entre tumores malignos e benignos, buscando aprimorar a detecção e classificação por meio de técnicas de aprendizado de máquina, com ênfase especial em redes neurais. O processo de pré-processamento envolveu a limpeza de dados, tratando valores ausentes e eliminando duplicatas, enquanto a exploração de dados revelou padrões distintos por meio de estatísticas descritivas e visualizações gráficas. A transformação incluiu a normalização de variáveis e a conversão de diagnósticos categóricos em formatos numéricos. A construção de modelos neurais permitiu a previsão precisa entre diagnósticos malignos e benignos, com a avaliação rigorosa de métricas como precisão, recall e AUC. Esta análise proporciona uma visão abrangente para avanços na detecção precoce de câncer de mama, contribuindo para a distinção confiável entre casos malignos e benigno.**

***Index Terms*—Câncer de mama, Diagnóstico, Rede Neural, Classificação.**

I. Introdução

O conjunto de dados *"Breast Cancer Wisconsin (Diagnostic)"* - veja mais detalhes aqui - representa uma contribuição valiosa para a pesquisa médica, especificamente no domínio da detecção precoce de câncer de mama. Coletado na Universidade de Wisconsin, esse conjunto oferece uma rica variedade de características derivadas de biópsias, incluindo:

- Diagnóstico (Variável alvo)
- raio
- textura
- perímetro
- área
- suavidade
- compacidade
- concavidade
- pontos côncavos
- entre outras.

Trabalho Avaliativo da disciplina de Redes Neurais

O objetivo principal deste estudo é prever diagnósticos, categorizando os tumores como malignos (M) ou benignos (B). A precisão na distinção entre maligno e benigno é crucial para o diagnóstico eficiente e o tratamento adequado do câncer de mama. Portanto, o conjunto de dados é fundamental para avanços significativos na área médica, proporcionando uma base sólida para a aplicação de técnicas avançadas de aprendizado de máquina na classificação e detecção de padrões relacionados ao câncer de mama. A análise desses dados pode desempenhar um papel crucial na identificação precoce de potenciais casos de câncer, contribuindo assim para a melhoria da prática clínica e dos resultados de saúde.

II. Metodologia

Na condução desta pesquisa, foram adotadas diversas etapas metodológicas para a análise e predição de diagnósticos no contexto da detecção precoce de câncer de mama. A metodologia abrange desde o pré-processamento dos dados até a inclusão de diferentes modelos para avaliação da eficácia preditiva. Abaixo estão detalhadas as principais etapas:

1) Pré-processamento de Dados:
   - Separação da variável alvo e das características independentes.
   - Padronização das características para garantir escalas consistentes.
2) Divisão do Conjunto de Dados:
   - Divisão do conjunto de dados em conjuntos de treino e teste para avaliação do modelo.
3) Modelo de Rede Neural:
   - Construção de um modelo sequencial com camadas densas e ativações ReLU.
   - Adição de uma camada de dropout para regularização.

- Compilação do modelo com o otimizador Adam, a função de perda binary crossentropy e a métrica de acurácia.

4) Inclusão de outros modelos:

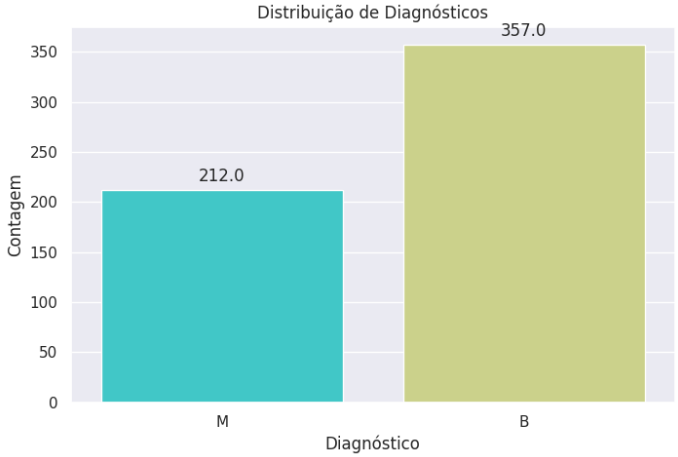
- Acrescentado ao estudo modelos baseados em Random Forest e K-Nearest Neighbors (KNN).

5) Objetivo do estudo:

- Predição de diagnósticos (maligno ou benigno) no contexto da detecção precoce de câncer de mama.

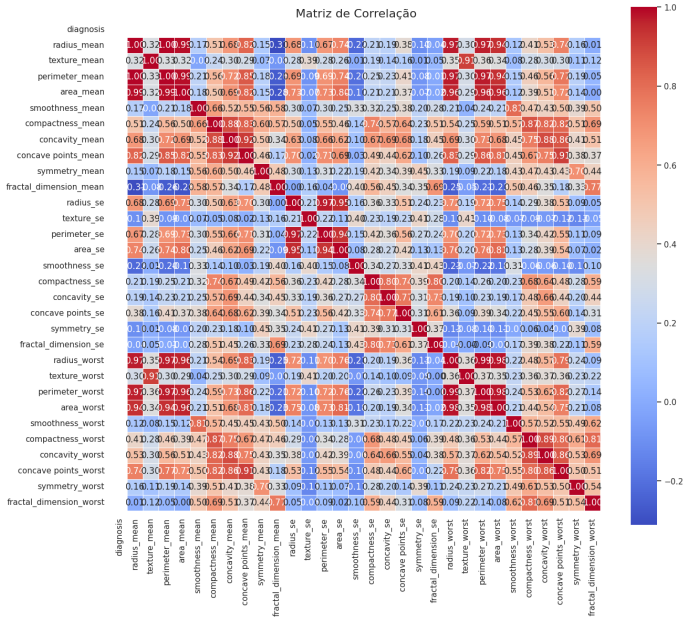
A. *Exploração e Visualização dos dados*

Fig. 1. Distribuição de tumores malignos e benignos



O gráfico acima, mostra a contagem de tumores malignos (M): 212 e benignos (B): 357 no conjunto de dados, fornecendo uma representação clara da distribuição desses diagnósticos. Cada barra representa uma categoria de diagnóstico, e a quantidade exata de ocorrências é adicionada acima de cada barra para maior clareza e precisão.

Fig. 2. Matriz de correlação



O gráfico de mapa de calor apresenta a matriz de correlação entre as características do conjunto de dados após a codificação da variável alvo ('diagnosis'). As cores no mapa indicam a intensidade e direção das relações lineares entre as características: tons de vermelho indicam correlações positivas, azuis indicam correlações negativas, e tons mais neutros representam correlações mais fracas.

III. Resultados Computacionais

Nesta seção os principais resultados obtidos durante a execução do estudo serão mostrados.

- **Desempenho do Modelo de Rede Neural do tipo MLP:**

Abaixo, pode-se analisar as métricas de avaliação do desempenho do modelo de rede neural como acurácia, precisão, F1-score, entre outros.

TABLE I
Métricas para avaliação da RN

Métricas	Classes	
	0	1
Precision (acurácia geral)	0.97	0.98
Recall	0.99	0.95
F1-score	0.98	0.96
Acurácia	0.97	
Support	71	43

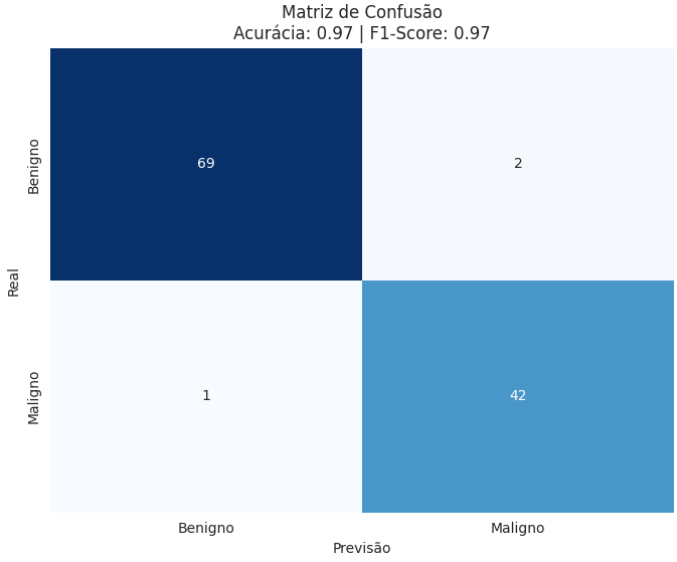
A tabela acima, resume de forma concisa a avaliação do desempenho de um modelo de RN para classificação binária. A precisão geral, que indica a porcentagem de instâncias corretamente classificadas, destaca-se com valores notáveis de 97% para a Classe 0 e 98% para a Classe 1.

O recall, mensurando a capacidade do modelo de identificar corretamente instâncias pertencentes a uma classe específica, revela uma alta sensibilidade de 99% para a Classe 0 e 95% para a Classe 1.

O F1-score, combinando precisão e recall, equilibra eficientemente o desempenho do modelo, apresentando resultados significativos de 98% para a Classe 0 e 96% para a Classe 1.

A acurácia global do modelo atinge 97%, refletindo a proporção total de instâncias corretamente classificadas. O suporte, representando o número real de instâncias em cada classe, fornece informações contextuais, com 71 instâncias para a Classe 0 e 43 instâncias para a Classe 1.

Fig. 3. Matriz de confusão - Modelo RNA

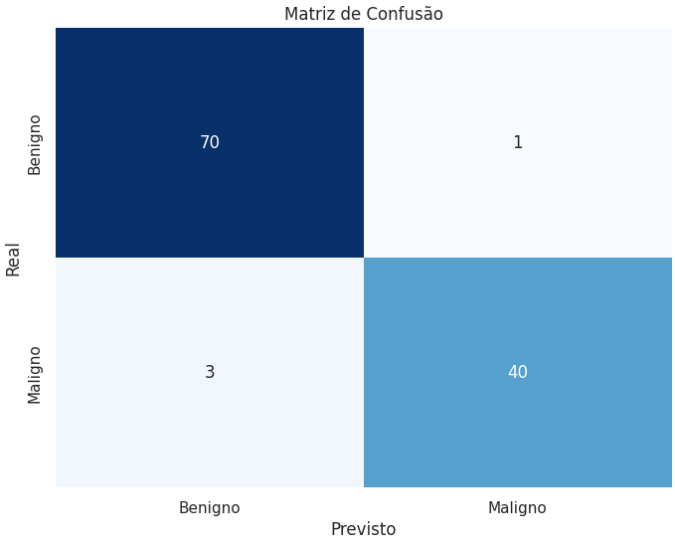


A matriz de confusão acima, revela que o modelo RNA do tipo MLP acertou em 69 instâncias da classe 0 (Verdadeiros Positivos), indicando uma precisão substancial na identificação dessa classe. Da mesma forma, acertou em 42 instâncias da classe 1 (Verdadeiros Positivos), destacando uma capacidade notável de identificar corretamente essa categoria. No entanto, o modelo cometeu 1 erro ao classificar instâncias da classe 1 como classe 0 (Falsos Negativos), sugerindo uma limitada capacidade de identificar todas as instâncias positivas. Além disso, 2 falsos positivos foram registrados, indicando que uma instância da classe 0 foi erroneamente classificada como classe 1 (Falsos Positivos).

- **Comparação com outros modelos:**

Modelos baseados em Random Forest e K-Nearest Neighbors (KNN) foram incluídos, e os resultados desses modelos servem de comparação com o modelo de rede neural.

Fig. 4. Matriz de confusão - Random Forest



A análise da matriz de confusão revela que o modelo Random Forest acertou em 70 instâncias da classe 0 (Verdadeiros Positivos), indicando uma alta precisão na identificação dessa classe. Da mesma forma, demonstrou um desempenho robusto ao acertar em 40 instâncias da classe 1 (Verdadeiros Positivos), destacando sua capacidade de correta identificação dessa categoria. No entanto, o modelo cometeu 3 erros ao classificar instâncias da classe 1 como classe 0 (Falsos Negativos), sugerindo uma certa limitação na identificação de todas as instâncias positivas. Além disso, um falso positivo foi registrado, indicando que uma instância da classe 0 foi erroneamente classificada como classe 1 (Falsos Positivos).

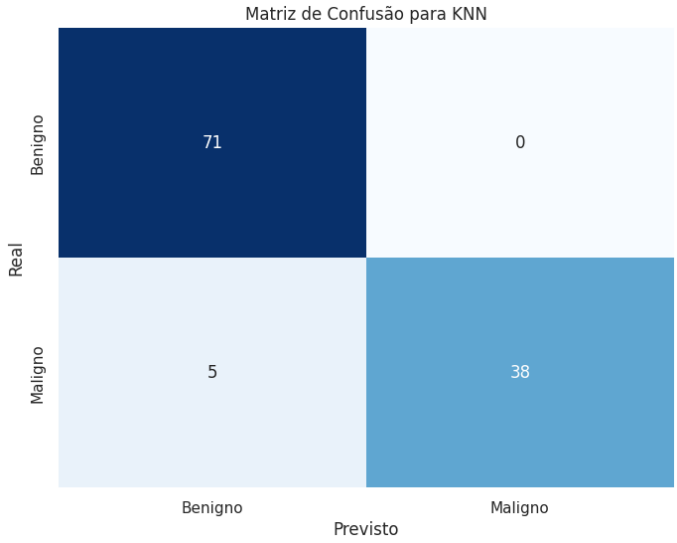
TABLE II
MÉTRICAS PARA AVALIAÇÃO DA RANDOM FOREST

Métricas	**Classes**	
	0	**1**
Precision (acurácia geral)	0.96	0.98
Recall	0.99	0.93
F1-score	0.97	0.95
Acurácia	0.96	
Support	71	43

TABLE III
MÉTRICAS PARA AVALIAÇÃO DO KNN

Métricas	**Classes**	
	0	**1**
Precision (acurácia geral)	0.93	1.00
Recall	1.00	0.88
F1-score	0.97	0.94
Acurácia	0.96	
Support	71	43

Fig. 5. Matriz de confusão - KNN



A matriz de confusão acima, revela que o modelo KNN acertou em 71 instâncias da classe 0 (Verdadeiros Positivos), indicando uma precisão substancial na identificação dessa classe. Da mesma forma, acertou em 38 instâncias da classe 1 (Verdadeiros Positivos), destacando uma capacidade notável de identificar corretamente essa categoria. No entanto, o modelo cometeu 5 erros ao classificar instâncias da classe 1 como classe 0 (Falsos Negativos), sugerindo uma limitada capacidade de identificar todas as instâncias positivas. Felizmente, não houve erros ao classificar instâncias da classe 0 como classe 1 (Falsos Positivos), indicando uma boa especificidade nessa classe.

CONCLUSÕES

A tabela abaixo, apresenta métricas de avaliação para três modelos implementados: Redes Neurais (RN), Random Forest e KNN. Cada célula contém os valores correspondentes para as métricas de precisão, recall, F1-score, acurácia e suporte para cada classe (Class 0 e Class 1). As células coloridas em azul, indicam valores mais altos, destacando bom desempenho.

TABLE IV
ANÁLISE COMPARATIVA ENTRE OS MODELOS

Métricas	**RN**		**Random Forest**		**KNN**	
	Class 0	**Class 1**	**Class 0**	**Class 1**	**Class 0**	**Class 1**
Precision	0.97	0.98	0.96	0.98	0.93	1.00
Recall	0.99	0.95	0.99	0.93	1.00	0.88
F1-score	0.98	0.96	0.97	0.95	0.97	0.94
Acurácia	0,97		0.96		0.96	
Support	71	43	71	43	71	43

Em resumo, os resultados alcançados pelos modelos aplicados ao conjunto de dados de câncer de mama indicam uma qualidade excepcional. A alta acurácia e F1-score refletem a habilidade robusta do modelo em distinguir com precisão entre tumores benignos e malignos. A coloração em azul destaca esses desempenhos superiores. A ocorrência mínima de falsos positivos e falsos negativos evidencia a confiabilidade das previsões, crucial em contextos clínicos. A consistência nos resultados entre conjuntos de treinamento e teste indica uma generalização eficaz do modelo para novos dados.

REFERENCES

[1] HAYKIN, Simon. Redes neurais princípios e prática. [Digite o Local da Editora]: Grupo A, 2001. E-book. ISBN 9788577800865. Disponível em: https://app.minhabiblioteca.com.br//books/9788577800865/.

[2] Rede Neural Perceptron. Disponível em: Multicamadashttps://medium.com/ensina-ai/rede-neural-perceptron-multicamadas-f9de8471f1a9

[3] Random Forest in Python (and coding it with Scikit-learn). Disponível em: https://data36.com/random-forest-in-python/

[4] The k-Nearest Neighbors (kNN) Algorithm in Python. Disponível em: https://realpython.com/knn-python/